



(12) 发明专利申请

(10) 申请公布号 CN 102253982 A

(43) 申请公布日 2011.11.23

(21) 申请号 201110172766.4

(22) 申请日 2011.06.24

(71) 申请人 北京理工大学

地址 100081 北京市海淀区中关村南大街 5
号

(72) 发明人 彭学平 牛振东 黄胜

(51) Int. Cl.

G06F 17/30 (2006.01)

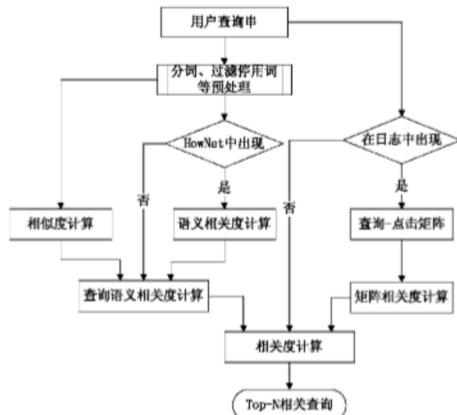
权利要求书 2 页 说明书 7 页 附图 2 页

(54) 发明名称

一种基于查询语义和点击流数据的查询建议
方法

(57) 摘要

本发明涉及一种基于查询语义和点击流数据的查询建议方法，包括以下步骤：一、对收集的查询日志数据进行预处理；二、对用户输入的查询数据进行分词、过滤停用词的预处理；三、将用户查询数据串与查询日志库中日志信息逐条进行相似度计算；四、基于知网中的词概念相关度计算方法，将用户查询数据串与查询日志库中日志信息逐条进行语义相关度计算；五、将相似度和语义相关度进行融合，计算用户查询数据串与查询日志库中每条日志信息的查询语义相关度；六、按照步骤五中的相关度由大到小，取出 Top-N 推荐给用户。本发明可以有效的消除查询歧义，并对输入错误进行提醒，提高信息检索系统的易用性和交互能力。



1. 一种基于查询语义和点击流数据的查询建议方法,包括以下步骤:

一、对收集的查询日志数据进行预处理,去掉非中文查询串、乱码数据及无意义的符号,形成规范的查询日志库;

二、对用户输入的查询数据进行分词、过滤停用词的预处理,形成包含多个关键词的查询数据串;

三、将用户查询数据串与查询日志库中日志信息逐条进行相似度计算;

四、基于知网中的词概念相关度计算方法,将用户查询数据串与查询日志库中日志信息逐条进行语义相关度计算;

五、将步骤三和步骤四计算出的相似度和语义相关度进行融合,计算用户查询数据串与查询日志库中每条日志信息的查询语义相关度;

六、按照步骤五中的相关度由大到小,取出 Top-N 推荐给用户。

2. 根据权利要求 1 所述的基于查询语义和点击流数据的查询建议方法,其特征在于,在得到用户查询数据串与查询日志库中每条日志信息的查询语义相关度之后,判断查询日志库中是否包含用户查询数据串,若不包含,则将用户查询数据串的矩阵相关度设为 0;若包含,则以用户提交的查询数据与该数据对应的点击 URL 之间的关系为基础,逐条计算用户查询数据串与查询日志库中其他查询日志信息之间的矩阵相关度;

将查询语义相关度和矩阵相关度进行融合,计算查询数据与查询日志库中每条日志信息的相关度,作为推荐给用户的依据。

3. 根据权利要求 1 或 2 所述的基于查询语义和点击流数据的查询建议方法,其特征在于,所述语义相关度计算方法为:

将用户查询数据串以及查询日志库中的每条日志信息均表示为规范化向量 $V(q) = (t_1, w_1; t_2, w_2; \dots; t_n, w_n)$, 其中 t_i 为特征项, w_i 为 t_i 在 q 中的权值;查询向量 $V(q)$ 中的每个元素的权值 w_i 由下面公式来计算,

$$w_i = \frac{freq_i}{\max \{ freq_j \mid j = (1, 2, \dots, n) \}}$$

其中, $freq_i$ 表示查询特征项 t_i 在查询 q 中的出现频率,而查询字符串 q 中总共包含 n 个特征项;

设用户查询数据串为 $V(q_1) = (t_1, w_1; t_2, w_2; \dots; t_n, w_n)$, 查询日志库中的一条日志信息为 $V(q_2) = (t_1, w_1; t_2, w_2; \dots; t_m, w_m)$, 则其语义相关度为:

$$ConcRel(q_1, q_2) = \sum_{i=1}^n \sum_{j=1}^m w_i \cdot w_j \cdot Sim(t_i, t_j)$$

其中 $i \in [1, n]$, $j \in [1, m]$, $Sim(t_i, t_j)$ 是知网定义的词之间的概念相似度;如果该词语不在知网的语义库中,则其概念相似度定义为 0。

4. 根据权利要求 1 或 2 所述的基于查询语义和点击流数据的查询建议方法,其特征在于,所述将相似度和语义相关度进行融合的方法为:

$$Sim(q_1, q_2) = \alpha \cdot SimKeywords(q_1, q_2) + (1 - \alpha) \cdot ConcRel(q_1, q_2)$$

其中 $SimKeywords(q_1, q_2)$ 是步骤三得到的相似度, $ConcRel(q_1, q_2)$ 是步骤四得到的语义相关度, α 是平衡系数,其取值范围在 $[0, 1]$ 范围内。

5. 根据权利要求 1 或 2 所述的基于查询语义和点击流数据的查询建议方法, 其特征在于, 所述矩阵相关度计算方法为:

(1) 构建一个二步图 $Bq1 = (Vq1, Eq1)$, 其中所有顶点集 $Vq1 = Q \cup L$, $Q = \{q_1, q_2, \dots, q_m\}$ 即用户提交查询的集合, $L = \{l_1, l_2, \dots, l_n\}$ 即用户点击的 URL 的集合; 所有边的集合 $Eq1 = \{(q_i, l_j) | \text{存在从 } q_i \text{ 到 } l_j \text{ 的一条边}\}$; 当且仅当一个用户提交了查询 q_i , 然后点击了 URL l_j , 边 (q_i, l_j) 存在;

把二步图 $Bq1$ 转换为一个矩阵 S , 对于 $m \times n$ 查询-URL 矩阵 S , 行表示查询, 列表示 URL, s_{ij} 的值表明一个查询 q_i 被不同用户连接到 $URLl_j$ 的次数, 这里的“不同”是指如果一个用户多次点击同一查询-URL 对, 只记为 1 次;

(2) 矩阵分解与相似度计算

定义优化函数如下:

$$\min_{Q,L} \|S - Q^T L\|_F^2 + \alpha \|Q\|_F^2 + \beta \|L\|_F^2$$

其中 α, β 为不大于 0.1 的正数, $\|\cdot\|_F$ 是弗罗宾尼范数, 最优化的目的是使两个规范化的低维矩阵乘积 $Q^T L$ 近似于 S ;

对上面公式做矩阵运算求解, 得到最优的 $d \times m$ 矩阵 Q , 矩阵的每一列是查询的 d 维特征向量; 向量的每个项用 w_{ij} 表示主成分, 其中 i 为列标, j 为行标, 且 $1 \leq i \leq m, 1 \leq j \leq d$; 两个查询的矩阵相关度采用空间余弦夹角进行计算, 其公式如下:

$$simMatrix(q_i, q_j) = \frac{\sum_{k=1}^d w_{i,k} \times w_{j,k}}{\sqrt{\sum_{k=1}^d w_{i,k}^2} \times \sqrt{\sum_{k=1}^d w_{j,k}^2}}$$

通过该公式计算得到两个查询的矩阵相关度。

6. 根据权利要求 1 或 2 所述的基于查询语义和点击流数据的查询建议方法, 其特征在于, 将查询语义相关度和矩阵相关度进行融合的方法为:

$$S_{(q,q_i)} = simMatrix(q, q_i) \cdot Sim(q, q_i)$$

其中 $S_{(q, q_i)}$ 为查询 q 和 q_i 融合基于查询语义和点击流矩阵的相关度。

7. 根据权利要求 6 所述的基于查询语义和点击流数据的查询建议方法, 其特征在于, 设定一个不大于 0.1 的正数, 当 $simMatrix(q, q_i) = 0$ 或 $Sim(q, q_i) = 0$ 时, 把这个正数赋值给 $simMatrix(q, q_i)$ 或 $Sim(q, q_i)$ 。

一种基于查询语义和点击流数据的查询建议方法

技术领域

[0001] 本发明涉及一种新的查询建议方法——基于查询语义和点击流数据的查询建议方法 QSQSCD (Query Suggestion Based on the Query Semantics and Click-through Data)，属于信息检索领域。

背景技术

[0002] 目前搜索引擎采用的主要交互方式是用户自主输入查询，搜索系统根据用户输入的查询提供检索结果。但是，很多时候用户输入的查询词并不能准确表达其搜索需求。一方面，用户输入的查询词通常比较短——平均只有两三个词；另一方面，很多搜索引擎含有歧义或意图模糊；此外，很多时候，用户之所以要使用搜索引擎进行信息的搜索就是因为对要检索话题知之甚少甚至毫无概念，这时候用户很难构造准确的查询。研究表明只有 25% 的查询能清晰表达用户的意图。

[0003] 为了更好地帮助用户构造查询，搜索引擎普遍采用查询建议技术，在搜索结果页面中的“相关搜索”就是查询建议的一个具体应用。查询建议指发现或构造一组与原查询 Q 相关的查询 {Q1, Q2, ...}，可以通过修改原查询 Q 或整个替换 Q 来实现这些相关查询。例如，对用户查询“苹果 iphone”，可以通过修改查询词“iphone”来推荐查询“苹果手机”，也可以将整个查询替换为“ipad”。

[0004] 由于有着巨大的应用需求和价值，查询建议成为近年来的研究热点。从技术实现上看，查询建议可以看作一个以搜索引擎查询为检索对象的信息检索问题。然而，不同于文档或网页，查询的自身特点使查询建议面临诸多挑战：

[0005] 首先，不同于文档或网页，查询通常只包含两到三个查询词，缺乏充分的文本内容，传统信息检索模型不适合直接对其进行处理；

[0006] 其次，用户查询信息稀疏。用户查询日志数据中多数查询出现次数很少，在对这些查询处理时，可利用的相关属性信息有限；

[0007] 最后，用户查询复杂多样。用户查询日志数据中通常包含几千万甚至上亿条不同的查询，即使是同一查询不同用户可能表示不同意图。此外，用户查询受时间、突发事件等因素影响。

[0008] 查询建议方法根据所依赖的数据不同可分为两类：基于文档的方法和基于日志的方法。
1) 第一种方法主要通过处理包含查询词的文档来分析查询，从相关文档或人工编辑语料中搜索找出与输入查询相关的词或短语，然后利用这些相关词或短语构建推荐查询。
2) 第二种方法主要通过分析用户的搜索引擎查询日志寻找曾经出现过的相似查询，然后向用户给予推荐。这两种方法各有利弊，基于日志的方法对处理出现频率小的稀疏查询比较困难，基于文档的方法虽能处理稀疏查询，但是查找相关文档也是一个难题。

发明内容

[0009] 本发明的目的是针对目前查询建议缺乏有效语义处理的问题，提出一种基于查询

语义和点击流数据的查询建议方法。

[0010] 本发明提供了一种基于查询语义和点击流数据的查询建议方法，包括以下步骤：

[0011] 一、对收集的查询日志数据进行预处理，去掉非中文查询串、乱码数据及无意义的符号，形成规范的查询日志库；

[0012] 二、对用户输入的查询数据进行分词、过滤停用词的预处理，形成包含多个关键词的查询数据串；

[0013] 三、将用户查询数据串与查询日志库中日志信息逐条进行相似度计算；

[0014] 四、基于知网中的词概念相关度计算方法，将用户查询数据串与查询日志库中日志信息逐条进行语义相关度计算；

[0015] 五、将步骤三和步骤四计算出的相似度和语义相关度进行融合，计算用户查询数据串与查询日志库中每条日志信息的查询语义相关度；

[0016] 六、按照步骤五中的相关度由大到小，取出 Top-N 推荐给用户。

[0017] 本发明还提出了基于点击流矩阵模型的矩阵相关度计算方法，并将其与查询语义相关度相融合，具体方法为：

[0018] 在得到用户查询数据串与查询日志库中每条日志信息的查询语义相关度之后，判断查询日志库中是否包含用户查询数据串，若不包含，则将用户查询数据串的矩阵相关度设为 0；若包含，则以用户提交的查询数据与该数据对应的点击 URL 之间的关系为基础，逐条计算用户查询数据串与查询日志库中其他查询日志信息之间的矩阵相关度；

[0019] 将查询语义相关度和矩阵相关度进行融合，计算查询数据与查询日志库中每条日志信息的相关度，作为推荐给用户的依据。

[0020] 有益效果

[0021] 本发明所述基于查询语义和点击流数据的查询建议方法，将查询语义信息以及查询数据与该数据对应的点击 URL 之间的关系作为查询建议的依据，可以有效的消除查询歧义，并对输入错误进行提醒，提高信息检索系统的易用性和交互能力。

附图说明

[0022] 附图 1. QSQSCD 的查询建议方法流程图；

[0023] 附图 2. 查询 - 点击二步图；

[0024] 附图 3. 查询建议平均精度比较。

具体实施方式

[0025] 下面结合附图，具体说明本发明的优选实施方式。

[0026] 本实施方式具体实现了本发明所述的基于查询语义和点击流数据的查询建议方法，其流程如图 1 所示，包括以下步骤：

[0027] 一、对收集的查询日志数据进行预处理，去掉非中文查询串、乱码数据及无意义的符号，形成规范的查询日志库；

[0028] 二、对用户输入的查询数据进行分词、过滤停用词的预处理，形成包含多个关键词的查询数据串；

[0029] 三、将用户查询数据串与查询日志库中日志信息逐条进行相似度计算；

[0030] 进行相似度计算可以使用多种方法,例如余弦相似度计算、皮尔森系数相似度计算等。此步骤是传统的文本相似度计算,通常基于词频统计计算文档相似度。但是如果仅仅只通过该步骤获得相似度,将会缺乏对文档语义的处理。如果相关文档之间的公共词较多,通过单纯基于词频的相似度计算方法可以达到相关计算的目的,如果相关文档之间的公共词较少,这种计算方法就难以取得较好的效果,特别对于较短的查询串。因为查询串中词汇的出现频率很小,如果把与之关联紧密的其他概念考虑进来,则可以凸现查询的语义。因此,本实施例在进行传统的相似度计算之后,在步骤四中进行语义相关度的计算。

[0031] 四、基于知网中的词概念相关度计算方法,将用户查询数据串与查询日志库中日志信息逐条进行语义相关度计算。

[0032] (1) 知网中的词概念相关度计算方法:

[0033] 知网中的每个词语均由 DEF 来描述其概念定义,DEF 的值由若干个义原以及它们与主干词之间的语义关系描述组成。知网中的概念是对词汇语义的描述,每个词的语义描述包含一个或多个概念,每个概念描述形成一个记录,概念的定义以及与之相关的同义、反义、上位、下位等关系,均描述于记录的 DEF 项中。比如 :DEF(高兴) = {aValue| 属性值, circumstances| 境况, happy| 福, desired| 良}。由于义原是 HowNet 中最小的语义单位,所以义原的相似度计算是概念相似度计算的基础。由于所有的义原根据上下位关系构成了一个树状的义原层次体系,所以采用简单的通过语义距离计算相似度的办法。假设两个义原在这个层次体系中的路径距离为 d, 两个义原 p_1, p_2 之间的语义距离为:

$$[0034] \quad Sim(p_1, p_2) = \frac{\alpha}{d + \alpha}$$

[0035] 其中, d 是 p_1 和 p_2 在义原层次体系中的路径长度,是一个正整数。 α 是一个可调节的参数,一般取经验值 $\alpha = 1.6$ 。

[0036] 知网中词语概念相似度计算的基本方法是通过计算部分之间的相似度得到整体的相似度。知网将一个词语概念的描述分成四个部分:

[0037] 1) 第一基本义原:其值为一个基本义原,我们将两个概念的这一部分的相似度记为 $Sim_1(S_1, S_2)$;

[0038] 2) 其它基本义原:对应于语义表达式中除第一基本义原描述式以外的所有基本义原描述式,其值为一个基本义原的集合,我们将两个概念的这一部分的相似度记为 $Sim_2(S_1, S_2)$;

[0039] 3) 关系义原:对应于语义表达式中所有的关系义原描述式,其值是一个特征结构,对于该特征结构的每一个特征,其属性是一个关系义原,其值是一个基本义原,或一个具体词。我们将两个概念的这一部分的相似度记为 $Sim_3(S_1, S_2)$;

[0040] 4) 关系符号:对应于语义表达式中所有的关系符号描述式,其值也是一个特征结构,对于该特征结构的每一个特征,其属性是一个关系义原,其值是一个集合,该集合的元素是一个基本义原,或一个具体词。我们将两个概念的这一部分的相似度记为 $Sim_4(S_1, S_2)$ 。

[0041] 于是,知网的词之间概念相似度由下式计算

$$[0042] \quad Sim(S_1, S_2) = \sum_{i=1}^4 \beta_i \prod_{j=1}^i Sim_j(S_1, S_2)$$

[0043] 其中, $\beta_i (1 \leq i \leq 4)$ 是可调节的参数,且有: $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$,

$\beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$ 。由于第一义原描述式反映了一个概念最主要的特征,所以一般将其权值定义得比较大,一般取在 0.5 以上。

[0044] (2) 语义相关度计算方法 :

[0045] 本发明提出的语义相关度是以知网中的词概念相关度为基础的。例如,可以直接计算两个查询串中每个词的概念相关度的加权和,来计算两个查询串的语义相关度;或者将两个查询串中概念相似度最大的两个词的概念相似度,作为两个查询串的语义相关度。总之要通过语义相关度的计算,将查询串之间的语义联系考虑进来,作为推荐给用户的一个重要依据。

[0046] 本实施例优选的语义相关度计算方法为 :

[0047] 将用户查询数据串以及查询日志库中的每条日志信息均表示为规范化向量 $V(q) = (t_1, w_1; t_2, w_2; L; t_n, w_n)$, 其中 t_i 为特征项, w_i 为 t_i 在 q 中的权值;查询向量 $V(q)$ 中的每个元素的权值 w_i 由下面公式来计算,

$$[0048] w_i = \frac{freq_i}{\max \{ freq_j \mid j = (1, 2, \dots, n) \}}$$

[0049] 其中, $freq_i$ 表示查询特征项 t_i 在查询 q 中的出现频率,而查询字符串 q 中总共包含 n 个特征项;

[0050] 设用户查询数据串为 $V(q_1) = (t_1, w_1; t_2, w_2; L; t_n, w_n)$, 查询日志库中的一条日志信息为 $V(q_2) = (t_1, w_1; t_2, w_2; L; t_m, w_m)$, 则其语义相关度为:

$$[0051] ConcRel(q_1, q_2) = \sum_{i=1}^n \sum_{j=1}^m w_i \cdot w_j \cdot Sim(t_i, t_j)$$

[0052] 其中 $i \in [1, n]$, $j \in [1, m]$, $Sim(t_i, t_j)$ 是知网定义的词之间的概念相似度;如果该词语不在知网的语义库中,则其概念相似度定义为 0;

[0053] 五、将步骤三和步骤四计算出的相似度和语义相关度进行融合,计算用户查询数据串与查询日志库中每条日志信息的查询语义相关度;本实施例中采用的融合方法为:

$$[0054] Sim(q_1, q_2) = \alpha \cdot SimKeywords(q_1, q_2) + (1 - \alpha) \cdot ConcRel(q_1, q_2)$$

[0055] 其中 $SimKeywords(q_1, q_2)$ 是步骤三得到的相似度, $ConcRel(q_1, q_2)$ 是步骤四得到的语义相关度, α 是平衡系数,其取值范围在 $[0, 1]$ 范围内。

[0056] 六、判断查询日志库中是否包含用户查询数据串,若不包含,则将用户查询数据串的矩阵相关度设为 0;若包含,则以用户提交的查询数据与该数据对应的点击 URL 之间的关系为基础,计算用户查询数据串与查询日志库中其他查询日志信息之间的矩阵相关度;

[0057] 点击流数据记录了 Web 用户的检索和点击活动,这些活动反映用户的兴趣及用户和查询、查询和点击文档之间的潜在语义关系。点击流数据的每一行包含下列信息:用户 ID(u), 用户提交的查询 (q), 用户点击的 URL(l), 点击的 URL 排序 (r), 查询提交的时间 (t), 如下表所示。

[0058]

| 访问时间 | 用户 ID | 查询 | 结果排名 | 点击顺序 | 点击的 URL |
|----------------------|----------------------|------|------|------|--|
| 20070703 18:35:36 | 2503070223 833044 | 西藏歌曲 | 1 | 1 | zhuying2006.blog.enorth.co m.cn/article/89382.shtml |
| 20070703 18:35:37 | 7667019683 329921 | 心理测试 | 2 | 2 | www.baihe.com/ |
| 20070703 18:35:38 | 2198145225 647654 | 大山诗歌 | 14 | 16 | novelchina.net/HTML/1064 2/282731.htm |
| 20070703 18:35:40 | 4746082363 665506 | 法国留学 | 4 | 1 | www.swliuxue.com/ |
| 20070703 18:35:51 | 9773355954 949936 | 江苏移动 | 3 | 2 | 211.138.198.10/main.jsp |
| ... | ... | ... | ... | ... | ... |

[0059] 因此点击流数据可以表示为 (u, q, l, r, t) 五元组集合。从统计学的观点来看, 对应一个网页的查询词集包含人对网页和提交查询之间的关系认知。因此, 本发明基于用户提交的查询数据与该数据对应的点击 URL 之间的关系, 定义了矩阵相关性, 作为为用户提供查询建议的一个重要依据。例如, 可以直接为对应相同网页的查询串设置一个非常大的矩阵相关性值, 或者直接计算两个查询串对应相同网页的个数, 并将该数值设置为矩阵相关性值。本实施例采取的矩阵相关度计算方法为:

[0060] (1) 构建一个二步图 $B_{q1} = (V_{q1}, E_{q1})$, 其中所有顶点集 $V_{q1} = Q \cup L$, $Q = \{q_1, q_2, \dots, q_m\}$ 即用户提交查询的集合, $L = \{l_1, l_2, \dots, l_n\}$ 即用户点击的 URL 的集合; 所有边的集合 $E_{q1} = \{(q_i, l_j) \mid \text{存在从 } q_i \text{ 到 } l_j \text{ 的一条边}\}$; 当且仅当一个用户提交了查询 q_i , 然后点击了 URL l_j , 边 (q_i, l_j) 存在;

[0061] 为了方便对 B_{q1} 执行矩阵降维和分解, 把二步图 B_{q1} 转换为一个矩阵 S , 对于 $m \times n$ 查询-URL 矩阵 S , 行表示查询, 列表示 URL, s_{ij} 的值表明一个查询 q_i 被不同用户连接到 URL_{l_j} 的次数, 这里的“不同”是指如果一个用户多次点击同一查询-URL 对, 只记为 1 次。这样能够较好的发现查询和 URL 之间的关系, 如图 2 所示。

[0062] (2) 矩阵分解与相似度计算

[0063] 对于 m 和 n 都达到千万级的时候, 矩阵 S 非常的庞大, 同时查询在二步图 B_{q1} 中是很稀疏的。比如, 在我们的实验数据中, 一个查询连接到平均 4.04 个 URL 上, 而且, 一个 URL 也仅涉及到很少的查询。在我们的实验中 URL 顶点的平均度只有 1.22。

[0064] 基于对查询-链接矩阵 S 的分析, 可以通过 S 的矩阵分解得到高质量低维度的查询 Q 和链接 L 的特征向量表示。新的特征表示提取了查询和链接的主要成分, 对进一步的处理更加有效。这里 Q 是一个 $d \times m$ 的矩阵, 每一列是查询的 d 维特征向量, 同时 L 是一个 $d \times n$ 矩阵, 每一列是链接的 d 维特征向量。

[0065] 我们可以使用类似于潜在语义索引 (LSI) 的方法, 应用著名的主成分分析 (PCA)

来得到 Q 和 L, 我们定义优化函数如下 :

$$[0066] \quad \min_{Q,L} \|S - Q^T L\|_F^2 + \alpha \|Q\|_F^2 + \beta \|L\|_F^2$$

[0067] 其中 α, β 为不大于 0.1 的正数, $\|\cdot\|_F$ 是弗罗宾尼范数 (Frobenius norm), 最优化的目的是使两个规范化的低维矩阵乘积 $Q^T L$ 近似于 S;

[0068] 根据对上面公式做矩阵运算求解, 得到最优的 $d \times m$ 矩阵 Q, 矩阵的每一列是查询的 d 维特征向量; 向量的每个项用 w_{ij} 表示主成分, 其中 i 为列标, j 为行标, 且 $1 \leq i \leq m$, $1 \leq j \leq d$; 两个查询的矩阵相关度采用空间余弦夹角进行计算, 其公式如下 :

$$[0069] \quad simMatrix(q_i, q_j) = \frac{\sum_{k=1}^d w_{i,k} \times w_{j,k}}{\sqrt{\sum_{k=1}^d w_{i,k}^2} \times \sqrt{\sum_{k=1}^d w_{j,k}^2}}$$

[0070] 七、将查询语义相关度和矩阵相关度进行融合, 计算查询数据与查询日志库中每条日志信息的相关度, 作为推荐给用户的依据。

[0071] 本实施方式中采用将查询语义相关度和矩阵相关度直接相乘的融合方法 :

$$[0072] \quad S_{(q, q_i)} = simMatrix(q, q_i) \cdot Sim(q, q_i)$$

[0073] 其中 $S_{(q, q_i)}$ 为查询 q 和 q_i 融合基于查询语义和点击流矩阵的相关度。但考虑到 $simMatrix(q, q_i)$ 和 $Sim(q, q_i)$ 中一个或两个可能等于 0。我们设定一个不大于 0.1 的正数, 比如为 0.01, 使得当 $simMatrix(q, q_i) = 0$ 或 $Sim(q, q_i) = 0$ 时, 把这个较小的正数赋值给 $simMatrix(q, q_i)$ 或 $Sim(q, q_i)$, 这样可以对模型做一个简单的平滑, 不至于出现零值。

[0074] 八、按照步骤七中的相关度由大到小, 取出 Top-N 推荐给用户。

[0075] 下表针对三组查询测试串 :“教育”、“旅游”和“健身”, 对本实施方式采用的查询建议方法 (QSQSCD) 与 Google、百度的“相关搜索”功能提供的查询建议进行比较。

[0076]

| 测试 查询 | 查询建议 | | | | |
|----------|------|------|------|------|------|
| | Top1 | Top2 | Top3 | Top4 | Top5 |

QSQSCD

| | | | | | |
|----|------|------|------|------|--------|
| 教育 | 教育部 | 教育网 | 职业教育 | 考试 | 培训 |
| 旅游 | 旅游保险 | 旅游同伴 | 驴友 | 旅游网 | 宾馆 |
| 健身 | 健身杂志 | 家庭健身 | 健身舞 | 健身器械 | 妇女健康杂志 |

Google

| | | | | | |
|----|------|---------|-------|---------|-------|
| 教育 | 教育新闻 | 吉林教育信息网 | 什么是教育 | 山东教师教育网 | 幼儿教育 |
| 旅游 | 旅游论坛 | 旅游网 | 旅游卫视 | 上海旅游 | 旅游线路 |
| 健身 | 健身方法 | 瑜伽健身 | 健身视频 | 中老年健身舞 | 广场健身舞 |

百度

| | | | | | |
|----|------|---------|------|--------|-------|
| 教育 | 教育部 | 人民教育出版社 | 教育网 | 中公教育 | 医学教育网 |
| 旅游 | 旅游网 | 五一旅游 | 北京旅游 | 西安旅游 | 张家界旅游 |
| 健身 | 健身器材 | 健身舞 | 健身房 | 郑多燕健身舞 | 健身计划 |

[0077] 在 Google、百度的“相关搜索”中均包含被测试的查询词，是对查询词进行查询扩展而得到的查询建议结果，不包含查询词的语义关系。而本发明提出的查询建议结果能反映查询词的相关语义信息，如用户查询“教育”在查询建议结果中会出现“考试”和“培训”相关词语，该词语能反映“教育”的语义信息，给用户有更深层次的提示和引导。在用户检索“旅游”时 QSQSCD 的查询建议结果中列出“驴友”、“宾馆”，经过分析发现是用户在搜索“旅游”和“驴友”时，有很多相同的点击 URL，同时“旅游”与用户的住宿存在语义关系，故“宾馆”被作为查询建议列举出来。

[0078] 在本实验中将本发明提出的查询建议方法 QSQSCD 和 SimRank 相似度计算方法进行了比较。SimRank 是利用图的结构信息计算对象间的相似度：一个节点与自身的相似度最高，相同或相似节点的邻居节点也相似。也就是说，节点间的相似性可以沿着边传递到他们的邻居间。下表展示的是对“教育”这个查询关键词在查询建议列表中次序为 1, 5, 10, 20 的查询建议精度。实验发现，本发明提出的查询建议方法在这四个位置的查询建议精度好于 SimRank 方法。

| [0079] | 推荐方法 | P@1 | P@5 | P@10 | P@20 |
|--------|---------|--------|--------|--------|--------|
| | QCQS | 0.8316 | 0.7505 | 0.6062 | 0.3062 |
| | SimRank | 0.8268 | 0.7228 | 0.5897 | 0.2597 |

[0080] 图 3 展示了 QSQSCD 和 SimRank 的平均查询建议精度，其中横坐标是位置 K 的值（从 1 到 10），纵坐标为在位置为 K 时的查询建议平均精度。在 K = 1 时，QSQSCD 和 SimRank 的平均查询建议精度都在 80% 以上，且非常的接近。但随着 K 的增多，也就是随着查询建议条目的增加，QSQSCDS 建议精度下降比 SimRank 更趋于平缓，前者的查询建议效果好于后者。

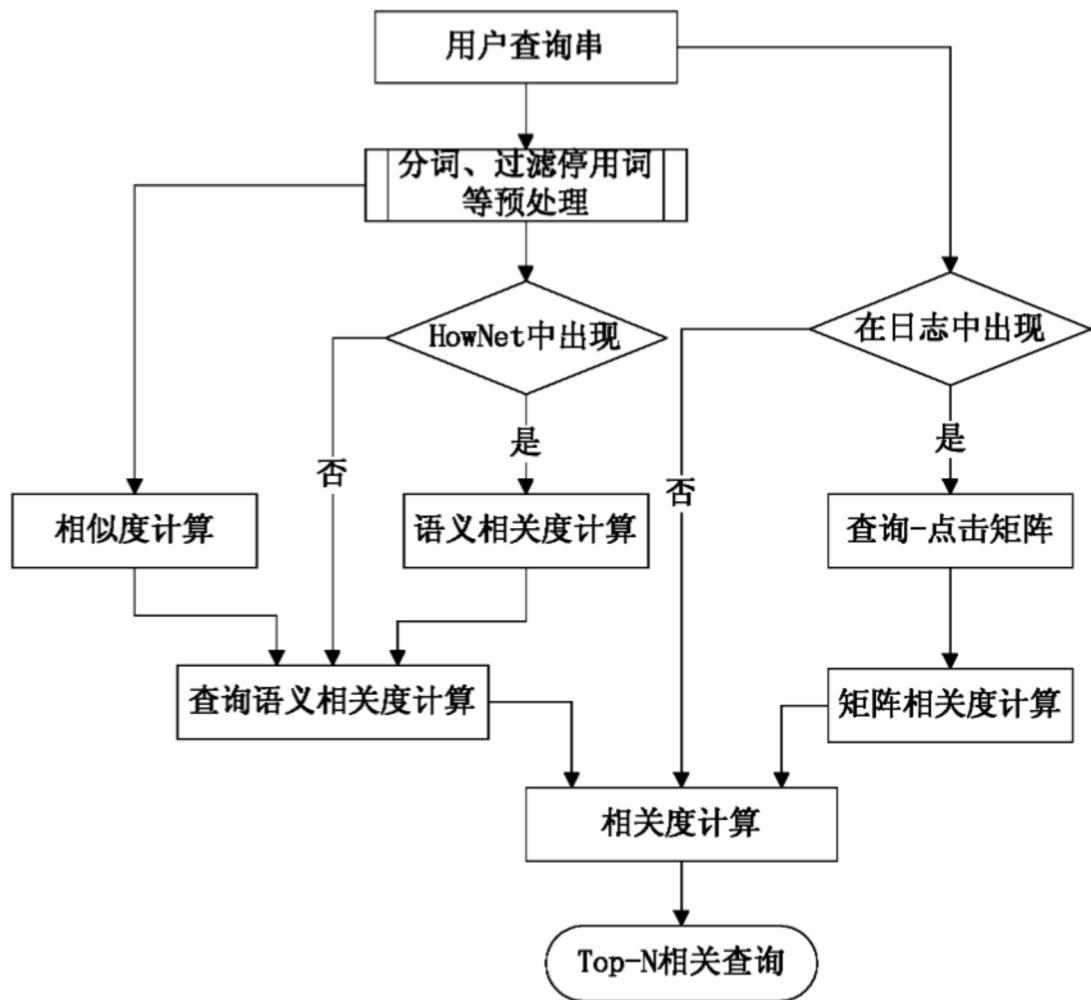


图 1

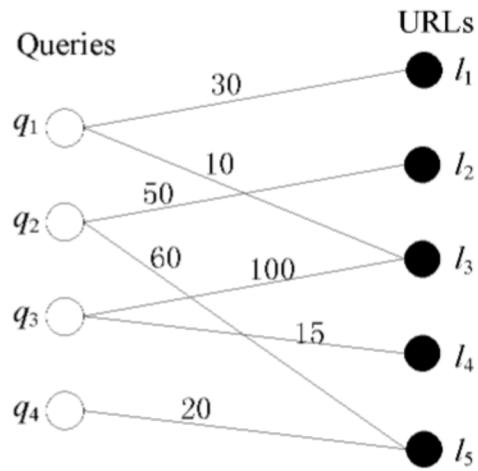


图 2

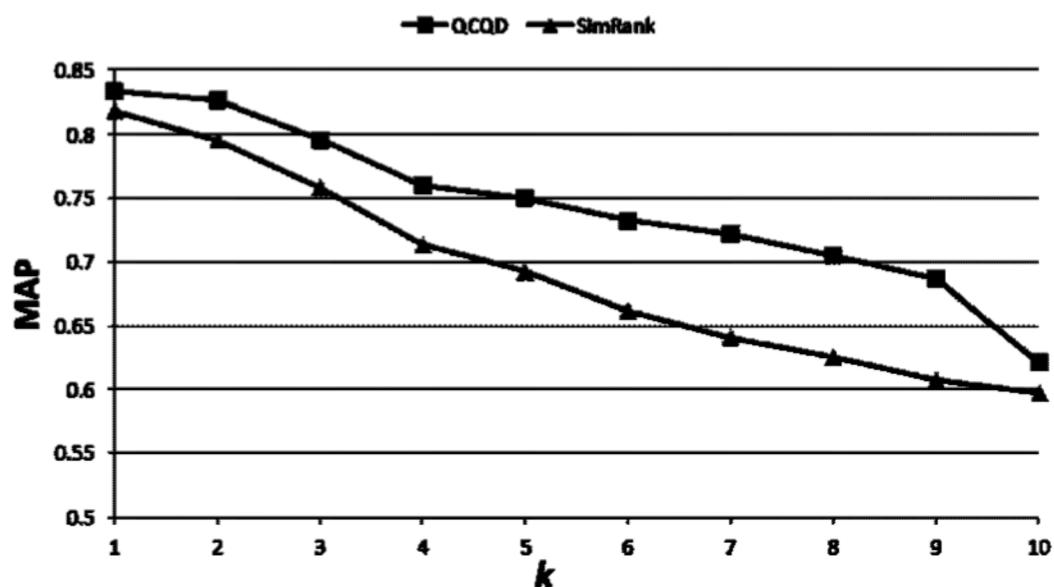


图 3